Fashion Retrieval with Contrastive Learning

Lê Hoàng Bùi
Faculty of Information Technology
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
0009-0003-1162-234X

Hoang Nguyen Cao
Faculty of Information Technology
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
0009-0003-0459-730X

Hoang-Nam Vo
Faculty of Information Technology
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
0009-0000-2835-3711

***Abstract*—**
***Index Terms*—**

I. Introduction

II. Related Work

Composed image retrieval (CIR) combines a query image with textual descriptions to retrieve target images, a challenging task requiring effective fusion of visual and textual modalities. Early approaches, such as Vo et al., 2019 [1], explored residual connection-based methods for integrating queries. Recent work leverages vision-language models: Liu et al., 2021 [2] introduced CIRPLANT with transformer-based adaptation, while Baldrati et al., 2022, 2023 [3], [4] utilized CLIP for feature fusion. Xu et al. [5] proposed Comquery-Former with a unified transformer architecture and global-local alignment. Zhao et al., 2022 [6] introduced progressive learning with adaptive weighting for hybrid queries, and Bai et al., 2023 [7] enhanced retrieval using sentence-level prompts from pretrained models. Other innovations include zero-shot methods (Saito et al. [8]), context-aware mapping (Tang et al. [9]), and application to video retrieval (Ventura et al. [10]).

Recent advances in **data generation for CIR** focus on synthetic data and multimodal retrieval techniques. CompoDiff [11], using latent diffusion, achieves state-of-the-art performance on benchmarks like Fashion IQ [12], and CIRR [2], through automated triplet generation. MagicLens [13] improves retrieval via self-supervised learning with synthesized embeddings, while InstructPix2Pix [14] generates text and images for compositional tasks using GPT-3 and diffusion models. Efforts like UniIR [15] integrate vision-language queries to enhance retrieval efficiency and accuracy across diverse datasets.

Negative sampling plays a vital role in contrastive learning for image retrieval. Feng et al. [16] used multi-modal language models to generate triplets for CIR, addressing positive data scarcity. Zhou and Li [17] proposed a coarse-to-fine alignment framework for cross-modal image retrieval, improving performance through targeted sampling. Contrastive learning approaches, such as SimCLR [18] and non-parametric instance discrimination [19], have shown that augmentations and effective sampling of negatives are essential for learning robust representations. Additionally, conditional negative sampling [20] enhances feature transferability to new distributions, while contrastive hashing with vision transformer [21] improves retrieval performance by integrating hard negative samples.

III. Methodology

A. Baselines

The baselines we used are derived from two key papers. Bai et al., 2023 [7] introduces the initial methodology for composed image retrieval (CIR) using sentence-level prompts. Feng et al., 2024 [16] presents improvements over the first by leveraging contrastive learning with scaling positives and negatives to enhance CIR performance. Figure 1 provides an overview of the methodology.

1) Preliminary: Suppose a CIR dataset consists of N annotated triplets consisting of reference images (r), modified text (t), and target image (t). A candidate image set consists of all reference images and target images. The task is to use reference image $r(i)$ and modified text $t(i)$ to compose a query $q(i)$. Then this $q(i)$ is used to retrieve target images from the candidate set.

2) Paradigm of Composed Image Retrieval: Multiple annotated triplets are combined into a mini-batch, and the reference images and modified texts in the same batch are then encoded using a query encoder to obtain query representations. The target images are encoded using an image encoder to obtain target image representations. The cosine similarity is then adopted to calculate the similarity between the query and target image representations.

3) Contrastive Learning: Treat the annotated examples as positive examples and treat the examples obtained by replacing the target image in the positive examples with another image in the mini-batch as the negatives. This approach pulls the query and target representations in positive examples closer while pushing the query and target representations in negative samples further. This yielded good results; however, the lack of positive examples and negative examples severely limits the full performance of contrastive learning, leading to optimization.

4) Scaling Positive Examples: Using multi-modal Large Language Model (MLLM) to construct the triplets for CIR.

a) Caption Generation: Design a prompt template to guide the MLLM to obtain a brief caption for each image under constrained conditions, where type and k are two dataset-specific parameters to simulate the type and length of modified text in the real dataset.

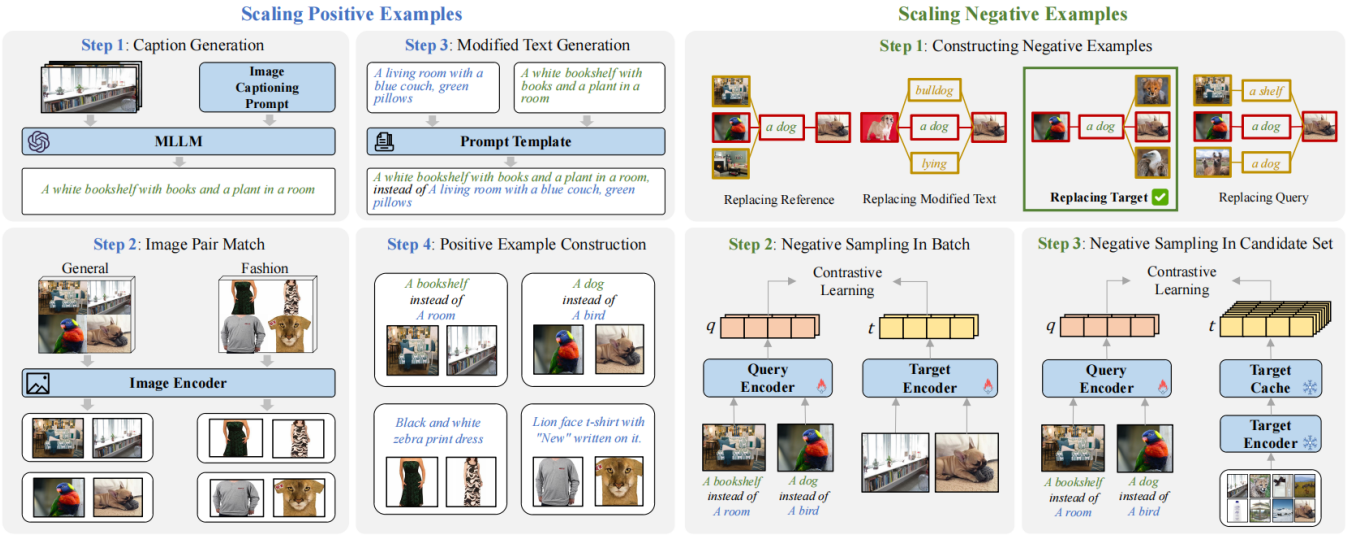

Figure 2: Overview of Our Framework of Scaling Positive Examples and Negative Examples. We abbreviate some of the modified texts due to space constraints.

Fig. 1. Overview of the methodology.

b) Image Pair Matching: Match two image-text pairs to generate a quadruplet. Introducing a uni-modal image encoder to get the representation of every image and calculate the pairwise similarity between two different images. Then we can rank the similarities related to the reference image in descending order. Only one image whose similarity rank is between $[c_0, c_1)$ $(c_0 < c_1)$ will be chosen as the target image, where c_0 and c_1 are two hyperparameters. This forms quadruplets.

c) Modified Text Generation: From the quadruplets, design a prompt template that generates modified text by either one of these:

- P_{temp0} : $\{C_t^i\}$ instead of $\{C^i\}$
- P_{temp1} : Unlike $\{C^i\}$, I want $\{C_t^i\}$
- P_{temp2} : $\{C_t^i\}$

d) Positive Example Construction: Combine image pairs from the second step with the modified text obtained in the third step to get new M triplets.

5) Scaling Negative Examples: Construct negative examples by replacing any element in the triplet. There are four methods to construct negative examples, and the authors examined each method to conclude that replacing the target image will obtain the best negative examples.

6) Two-stage Finetuning: In the first stage, fine-tune both the query encoder and the target encoder with in-batch negative sampling. In the second stage, freeze the target encoder and only fine-tune the query encoder.

B. Potential Approaches

We aim to improve the baseline model by integrating new techniques and optimizations.

1) Image Generation: One potential approach is to generate images that seamlessly align with the reference image and relative captions. By integrating a new image, we think that the model can incorporate image-side information to improve the performance of the model. We propose to input the reference image and text into the Multi-Instance Generation Controller (MIGC) [22] to generate a new image. This new image can provide useful information that are difficult to express in text at a fine-grained level.

2) Two-stage Optimization: We also propose a two-stage approach to optimize the performance. First, we filter out the candidate images that are not relevant to the reference image and modified text. Then, we use the filtered candidate images to retrieve the target image. This approach can reduce the search space and improve the retrieval performance.

IV. Experiments

We are currently conducting experiments to replicate the results of the original papers and to evaluate the impact of various learning techniques on retrieval accuracy. We intend to conduct additional experiments in the future following the implementation of certain improvements.

A. Datasets

- **FashionIQ**: At present, we are utilizing only the *FashionIQ* dataset [12], as employed by the original study. FashionIQ is a natural language-based interactive fashion product retrieval dataset. It contains 77,684 images crawled from Amazon.com, covering three categories: Dresses, Tops & Tees, and Shirts. Among the 46,609 training images, there are 18,000 image pairs. Each pair is accompanied by an average of two natural language sentences that describe one or multiple visual properties

to modify in the reference image, such as "is shiny" or "is blue in color and floral, and with white base."

- **Fashion200K**: We anticipate incorporating the *Fashion200K* dataset [23] in subsequent experiments. Fashion200K is a large-scale fashion dataset crawled from multiple online shopping websites. It contains more than 200,000 fashion images collected for attribute-based product retrieval, covering five categories: Dresses, Jackets, Pants, Skirts, and Tops. Each image is tagged with descriptive text corresponding to a product description, such as "beige v-neck bell-sleeve top."
- **Shopping100K**: We also plan to use the *Shopping100K* dataset [24] in future experiments. Shopping100K is a large-scale fashion dataset of individual clothing items extracted from different e-commerce providers. It contains 101,021 images of 12 fashion attributes, covering categories such as "collar," "color," "fabric," "fastening," "fit," "gender," "length," "neckline," "pattern," "pocket," "sleeve length," and "sport." Compared to FashionIQ and Fashion200K, the Shopping100K dataset is more diverse and only contains garments in isolation.

B. Evaluation Metrics

We employed *Recall@K (R@K)* [25] as the primary evaluation metric, which measures the proportion of queries for which the retrieved top K images include the correct target image. *Recallsubset@K (Rsubset@K)* is a similar metric, but the model retrieves images only within the semantically similar group of the reference image.

C. Current Results

We obtained results that are highly consistent with those reported in the original paper. Figure 2 shows the original results.

To further demonstrate the performance of our model, we present four images illustrating success and failure cases (see Figure 3).

V. Conclusion

References

[1] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval - an empirical odyssey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[2] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 2125–2134.

[3] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Conditioned and composed image retrieval combining and partially fine-tuning clip-based features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2022, pp. 4959–4968.

[4] A. Baldrati, M. Bertini, T. Uricchio, and A. del Bimbo, "Composed image retrieval using contrastive learning and task-oriented clip-based features," 2023. [Online]. Available: https://arxiv.org/abs/2308.11485

[5] Y. Xu, Y. Bin, J. Wei, Y. Yang, G. Wang, and H. T. Shen, "Multi-modal transformer with global-local alignment for composed query image retrieval," *Trans. Multi.*, vol. 25, no. 1, pp. 8346–8357, Jan. 2023. [Online]. Available: https://doi.org/10.1109/TMM.2023.3235495

[6] Y. Zhao, Y. Song, and Q. Jin, "Progressive learning for image retrieval with hybrid-modality queries," 2022. [Online]. Available: https://arxiv.org/abs/2204.11212

[7] Y. Bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R. S. M. Goh, and C.-M. Feng, "Sentence-level prompts benefit composed image retrieval," 2023. [Online]. Available: https://arxiv.org/abs/2310.05473

[8] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 19 305–19 314.

[9] Y. Tang, J. Yu, K. Gai, J. Zhuang, G. Xiong, Y. Hu, and Q. Wu, "Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5180–5188, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28324

[10] L. Ventura, A. Yang, C. Schmid, and G. Varol, "Covr-2: Automatic data construction for composed video retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 409–11 421, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2024.3463799

[11] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2303.11916

[12] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," 2020. [Online]. Available: https://arxiv.org/abs/1905.12794

[13] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, "Magiclens: Self-supervised image retrieval with open-ended instructions," 2024. [Online]. Available: https://arxiv.org/abs/2403.19651

[14] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," 2023. [Online]. Available: https://arxiv.org/abs/2211.09800

[15] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen, "Uniir: Training and benchmarking universal multimodal information retrievers," 2023. [Online]. Available: https://arxiv.org/abs/2311.17136

[16] Z. Feng, R. Zhang, and Z. Nie, "Improving composed image retrieval via contrastive learning with scaling positives and negatives," 2024. [Online]. Available: https://arxiv.org/abs/2404.11317

[17] L. Zhou and Y. Li, "Coarse-to-fine alignment makes better speech-image retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2408.13119

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[19] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," 2018. [Online]. Available: https://arxiv.org/abs/1805.01978

[20] M. Wu, M. Mosse, C. Zhuang, D. Yamins, and N. Goodman, "Conditional negative sampling for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2010.02037

[21] X. Ren, X. Zheng, H. Zhou, W. Liu, and X. Dong, "Contrastive hashing with vision transformer for image retrieval," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 12 192–12 211, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.23082

[22] D. Zhou, Y. Li, F. Ma, X. Zhang, and Y. Yang, "Migc: Multi-instance generation controller for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 6818–6828.

[23] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," 2017. [Online]. Available: https://arxiv.org/abs/1708.01311

[24] Y. Hou, E. Vig, M. Donoser, and L. Bazzani, "Learning attribute-driven disentangled representations for interactive fashion retrieval," in *The International Conference on Computer Vision (ICCV)*, October 2021.

[25] Y. Patel, G. Tolias, and J. Matas, "Recall@k surrogate loss with large batches and similarity mixup," 2022. [Online]. Available: https://arxiv.org/abs/2108.11179

Table 2: Evaluation results of various models on FashionIQ. The best results are in boldface.

Methods	Backbone	Dress		Shirt		Top&Tee		Average		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Rmean
CIRPLANT [21]	w/o VLP	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
ARTEMIS [9]	w/o VLP	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	38.17
ComqueryFormer [39]	w/o VLP	28.85	55.38	25.64	50.22	33.61	60.48	29.37	55.36	42.37
PL4CIR [41]	CLIP	33.60	58.90	39.45	61.78	43.96	68.33	39.02	63.00	51.01
TG-CIR [35]	CLIP	35.55	59.44	40.24	62.37	43.65	67.36	39.81	63.06	51.44
+SPN	CLIP	36.84	60.83	41.85	63.89	45.59	68.79	41.43	64.50	52.97
CLIP4CIR [4]	CLIP	38.18	62.67	44.01	64.57	45.39	69.56	42.52	65.60	54.06
+SPN	CLIP	**38.82**	**62.92**	**45.83**	**66.44**	**48.80**	**71.29**	**44.48**	**66.88**	**55.68**
BLIP4CIR [23]	BLIP	44.22	67.08	45.00	66.68	49.72	73.02	46.31	68.93	57.62
+SPN	BLIP	**44.52**	**67.13**	**45.68**	**67.96**	**50.74**	**73.79**	**46.98**	**69.63**	**58.30**
SPRC [2]	BLIP-2	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
+SPN	BLIP-2	**50.57**	**74.12**	**57.70**	**75.27**	**60.84**	**79.96**	**56.37**	**76.45**	**66.41**

Table 3: Performance comparison of various models on CIRR. The best results are in boldface.

Methods	Backbone	Recall@K				R_{subset}@K			Rmean
		K=1	K=5	K=10	K=50	K=1	K=2	K=3	
CIRPLANT [21]	w/o VLP	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS [9]	w/o VLP	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
ComqueryFormer [39]	w/o VLP	25.76	61.76	75.90	95.13	51.86	76.26	89.25	56.81
TG-CIR [35]	CLIP	45.23	78.34	87.13	97.30	72.84	89.25	95.13	75.59
+SPN	CLIP	**47.28**	**79.13**	**87.98**	97.54	**75.40**	**89.78**	95.21	**77.27**
CLIP4CIR [4]	CLIP	42.80	75.88	86.26	97.64	70.00	87.45	94.99	72.94
+SPN	CLIP	45.33	78.07	87.61	**98.17**	73.93	89.28	**95.61**	76.00
BLIP4CIR [23]	BLIP	44.77	76.55	86.41	**97.18**	74.99	89.90	95.59	75.77
+SPN	BLIP	**46.43**	**77.64**	**87.01**	97.06	**75.74**	**90.07**	**95.83**	**76.69**
SPRC [2]	BLIP-2	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
+SPN	BLIP-2	**55.06**	**83.83**	**90.87**	**98.29**	**81.54**	**92.65**	**97.04**	**82.69**

Fig. 2. The original paper's results.

Fig. 3. Illustrative examples of the model's performance. The first two images represent success cases, while the last two images represent failure cases.